DOI:10.16644/j.cnki.cn33-1094/tp.2025.06.007

植物可翻译lncRNA预测方法的比较研究

曹子雍[1,2]，刘长宁[1]

（1. 中国科学院西双版纳热带植物园，云南 西双版纳 666303；2. 中国科学院大学）

摘　要：随着植物长非编码RNA研究的深入，部分lncRNA（长非编码RNA）被发现能通过小开放阅读框（smORFs）翻译功能性小肽，这些小肽在植物生长、发育和逆境应答中发挥重要作用。本文主要比较了现有用于植物可翻译lncRNA识别的计算方法的差异与优势，包括基于序列特征、机器学习模型和多源数据融合的预测方法。通过对不同计算方法的评估，本文总结了它们在准确率、灵敏度、特异性等方面的优缺点，并提出了未来可能的改进方向，如深度学习的应用和多组学数据的结合。

关键词：植物lncRNA；可翻译lncRNA；计算方法；机器学习；序列特征；多组学数据

中图分类号：Q811.4　　**文献标识码：**A　　**文章编号：**1006-8228(2025)06-33-05

A Comparative Evaluation of Computational Methods for Identifying Translatable lncRNAs in Plants

Cao Ziyong[1,2], Liu Changning[1]

(*1. Xishuangbanna Tropical Botanical Garden Chinese Academy of Sciences, Xishuangbanna, Yunnan 666303, China; 2. University of Chinese Academy of Sciences*)

Abstract: With the advancement of plant long non-coding RNA (lncRNA) research, some lncRNAs have been found to translate functional micropeptides through small open reading frames (smORFs), which play crucial roles in plant growth, development, and stress responses. This paper primarily compares the differences and advantages of existing computational methods for identifying translatable lncRNAs in plants, including sequence-based approaches, machine learning models, and multi-source data integration methods. By evaluating different computational methods, this paper summarizes their strengths and weaknesses in terms of accuracy, sensitivity, and specificity, and proposes potential future improvements, such as the application of deep learning and the integration of multi-omics data.

Keywords: Plant lncRNA; Translatable lncRNA; Computational Methods; Machine Learning; Sequence Features; Multi-omics Data

0 引言

近年来，随着基因组与转录组技术的快速发展，长非编码RNA（lncRNA）被发现不仅在调控水平发挥作用，还可通过小开放阅读框（smORFs）翻译产生功能性小肽[1,2]。为了进一步分类和研究lncRNA，本研究将这类具备翻译潜力的lncRNA称作“可翻译lncRNA”（translatable lncRNA，简称tr-lncRNA）。这些tr-lncRNAs在植物生长发育、激素响应和逆境适应中展现出重要的生物学意义，逐渐成为植物功能基因组学研究的重要方向[3-7]。

传统识别tr-lncRNA的实验手段主要依赖Ribo-seq、质谱等高通量技术，虽具高精度但存在成本高、组织特异性强等局限，限制了其在大规模植物样本中的广泛应用。因此，构建基于序列信息的计算预测方法，成为识别tr-lncRNA的有效替代方案。目前，研究者已开发多种基于机器学习的预测模型，用于判断lncRNA是否具有翻译能力，并尝试将其应用于植物系统。

本文聚焦于两种代表性tr-lncRNA预测方法：基于深度神经网络的LncReader[8]与基于自组织映射和感知器组合的浅层学习模型IRSOM2[9]。前者整合序列、理化性质和二级结构信息，并采用多头自注意力

收稿日期：2025-05-21

作者简介：曹子雍（2000—），男，湖南衡阳人，硕士，主要研究方向：生物信息学。

通讯作者：刘长宁（1978—），男，湖南长沙人，博士，研究员，主要研究方向：lncRNA功能研究。

机制增强模型表达能力；后者则依赖多种ORF相关特征，并引入拒绝选项机制以提升预测可信度。虽然二者在动物系统中均取得了良好效果，但在植物系统中表现差异显著，尚缺乏系统的比较研究。

为此，本文选取拟南芥（*Arabidopsis thaliana*）与玉米（*Zea mays*）两个模式植物作为测试对象，构建标准植物tr-lncRNA数据集，并将现有模型LncReader与IRSOM2应用于该数据，评估其在植物体系中的表现。通过准确率、F1分数、AUC值及跨物种泛化能力等指标对比分析，揭示不同模型在跨物种迁移预测与识别精度方面的差异。研究结果有助于明确各模型在植物数据上的适用性与局限性，为今后tr-lncRNA计算方法在植物系统中的优化与选择提供理论依据。

1 方法功能与原理分析

本研究比较了两种代表性tr-lncRNA识别模型：LncReader和IRSOM2，它们分别代表了深度学习与浅层统计学习方法的最新进展。

1.1 LncReader：融合多模态特征的深度注意力网络

LncReader是一种基于多头自注意力机制的深度学习模型，旨在识别具备翻译能力的双功能lncRNA。该模型融合了三类特征输入：序列特征（如ORF长度、Fickett得分等）、理化性质（如等电点、电子—离子相互作用势）和RNA二级结构（如最小自由能）。为解决序列信息长程依赖问题，LncReader采用了多头自注意力机制（Multi-head Self Attention），在提升模型表达力的同时，有效降低了局部特征的过拟合风险。

在模型结构上，LncReader基于Transformer架构构建，使用多层并行编码器进行上下文建模，并通过交叉层参数共享机制减小模型复杂度。其最终分类模块由线性变换和Sigmoid函数输出组成，可直接输出概率得分用于判断双功能潜能。值得一提的是，该模型在结合RNA-seq、Ribo-seq与MS等多组学验证数据构建的数据集上表现出极高的准确性与稳定性，在五折交叉验证的基准测试中优于DNN、CNN、SVM等传统分类器。

然而需要注意的是，LncReader最初仅在人类数据上训练，尚未对植物tr-lncRNA特征进行定制优化，直接迁移至拟南芥和玉米等植物物种时，其预测性能可能受限于物种间lncRNA序列结构和翻译信号的系统差异。

1.2 IRSOM2：基于模糊拒绝的三分类浅层模型

IRSOM2是在原始IRSOM模型基础上发展而来的三分类预测工具，主要用于识别RNA的“编码”，“非编码”和“双功能”潜力。该方法本质上基于二元监督学习，但通过引入模糊性拒绝机制（Ambiguity-based Rejection）构建出三类判别体系。

在模型结构上，IRSOM2使用自组织映射网络（Self-Organizing Map, SOM）提取数据的拓扑表示，并在此基础上构建感知器层进行分类。输入特征包括：k-mer频率、ORF统计特征（最大ORF长度、覆盖率、覆盖率分布等）、核苷酸组成（GC含量、密码子位置偏差指数等）。

IRSOM2的关键在于：当模型对编码或非编码的分类置信度差异低于设定阈值（如0.8）时，将样本标记为“双功能RNA”或“拒绝分类”。该策略不仅避免了对边界样本的误判，还使得预测结果具备一定程度的可解释性，适合于模糊区域转录本的筛选与分析。

IRSOM2已预置了多个物种的训练模型，包括人类、拟南芥、水稻等，并允许用户在线上传FASTA序列进行快速分类，适合植物转录本的大规模筛查任务。不过，该工具的模型本质仍依赖浅层特征工程，缺乏对序列上下文与非线性依赖关系的建模能力，从而限制了对复杂结构tr-lncRNA的识别准确率。

2 数据来源与测试集构建

本研究旨在评估现有tr-lncRNA预测模型LncReader与IRSOM2在植物系统中的适用性与泛化能力。为此，选取两个模式植物拟南芥（*Arabidopsis thaliana*）与玉米（*Zea mays*），构建高质量tr-lncRNA正负样本集作为模型评估数据基础。

2.1 数据来源与预处理

质谱数据（MS）来源于公开植物小肽翻译数据库PsORF[10]，并结合Ali等[11]与Wang等[12]补充的翻译肽段数据。肽段与lncRNA转录本序列比对后，保留覆盖开放阅读框（ORF）区域、具有明确翻译证据的匹配项，作为MS支持的候选tr-lncRNA（MS-lncRNAs）。

核糖体足迹测序数据（Ribo-seq）来自NCBI SRA数据库中发布的11个高质量植物Ribo-seq项目[13-23]（如表1所示）。数据预处理包括接头和低质量碱基剪切（Trim Galore）、rRNA reads去除（Bowtie）、参考基因组比对（Hisat2）和排序统计（Samtools）。以lncRNA中ORF区域的核糖体富集为标准，识别Ribo-seq支持的转录本（Ribo-seq-lncRNAs）。

表1 拟南芥和玉米Ribo-seq数据来源

物种	Accession	来源
Arabidopsis thaliana	PRJNA173092	文献[13]
	PRJNA286795	文献[14]
	PRJNA321304	文献[15]
	PRJNA342301	文献[16]
	PRJNA328073	文献[17]
	PRJNA594648	文献[18]
	PRJNA854638	文献[19]
Zea mays	PRJNA272662	文献[20]
	PRJNA313100	文献[21]
	PRJNA435622	文献[22]
	PRJNA530618	文献[23]

2.2 测试集构建策略

为构建生物学合理的评估数据集，本研究采用多源证据筛选策略。正类样本包含Ribo-seq-lncRNAs、MS-lncRNAs及其交集（Ribo-MS-lncRNAs）。负类样本（untr-lncRNAs）通过从全部lncRNA集合中去除任何具有翻译证据的转录本（即Ribo-seq或MS支持的转录本）得到，确保其不含任何翻译活性证据。

最终，在两个物种中分别构建高置信度的正负样本集。本文主要选取Ribo-seq-lncRNAs与untr-lncRNAs作为测试集，用于LncReader与IRSOM2模型评估。各类lncRNA的数量统计如表2所示。

表2 各类lncRNA的数据集构建统计

物种	lncRNA类别	基因数量	转录本数量	平均转录本数量
Arabidopsis thaliana	Total lncRNA	13,664	16,537	1.21
	Ribo-seq-lncRNAs	1,876	2,157	1.15
	MS-lncRNAs	37	54	1.46
	untr-lncRNA	11,770	13,621	1.16
Zea mays	Total lncRNA	28,269	33,705	1.19
	Ribo-seq-lncRNAs	1,680	1,961	1.17
	MS-lncRNAs	52	60	1.15
	untr-lncRNA	26,548	30,965	1.17

3 模型评估与结果分析

为评估现有tr-lncRNA预测方法LncReader和IRSOM2在植物中的适用性，本文基于拟南芥（Arabidopsis thaliana）与玉米（Zea mays）两个代表性植物物种，系统比较了LncReader与IRSOM2两种模型的预测性能。从评估结果来看，两种方法在植物数据中均表现出不同程度的识别能力下降，这提示现有模型在跨物种迁移至植物系统时面临显著挑战。

3.1 ROC曲线与整体判别能力

图1展示了LncReader与IRSOM2在拟南芥（图1A）和玉米（图1B）测试集上的ROC曲线表现。结果显示，两种模型在两个物种中均具备一定的tr-lncRNA判别能力，整体表现较为接近。其中，LncReader在拟南芥和玉米中分别获得0.66的AUC值，IRSOM2则分别为0.59和0.62，表明LncReader在总体上略具优势，但差异幅度有限，尚不足以构成显著性能差异。

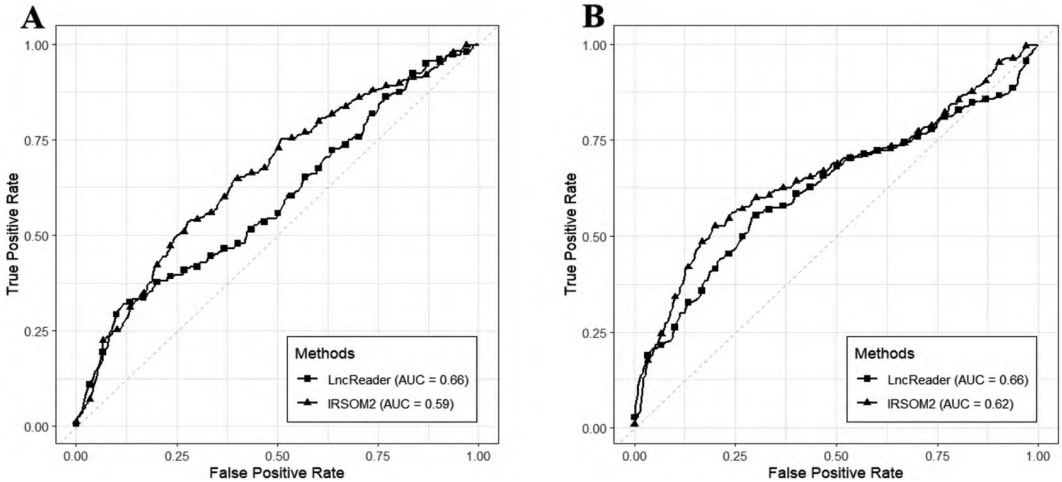

图1 LncReader与IRSOM2在拟南芥和玉米测试集上的ROC曲线比较

需要注意的是，尽管两个模型在AUC指标上均达到中等水平，但均未超过0.70的理想判别阈值，表明当前模型在植物系统中的tr-lncRNA识别仍面临一定挑战。植物lncRNA普遍具有较低序列保守性、多样化结构特征及非典型翻译信号，这些因素可能限制了现有模型在该类转录本上的判别能力。

综合来看，LncReader与IRSOM2在跨物种应用中判别性能相当，均未能展现出高水平的分类能力，提示现有模型在植物tr-lncRNA识别任务中仍有较大优化空间，亟需结合植物系统的特异性信息进行改进。

3.2 分类指标对比分析

为了更全面地评估模型性能，本研究进一步从准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1分数（F1-score）等维度，对LncReader与IRSOM2在两个植物物种中的分类表现进行了比较分析（如图2所示）。相关数值统计如表3所示。

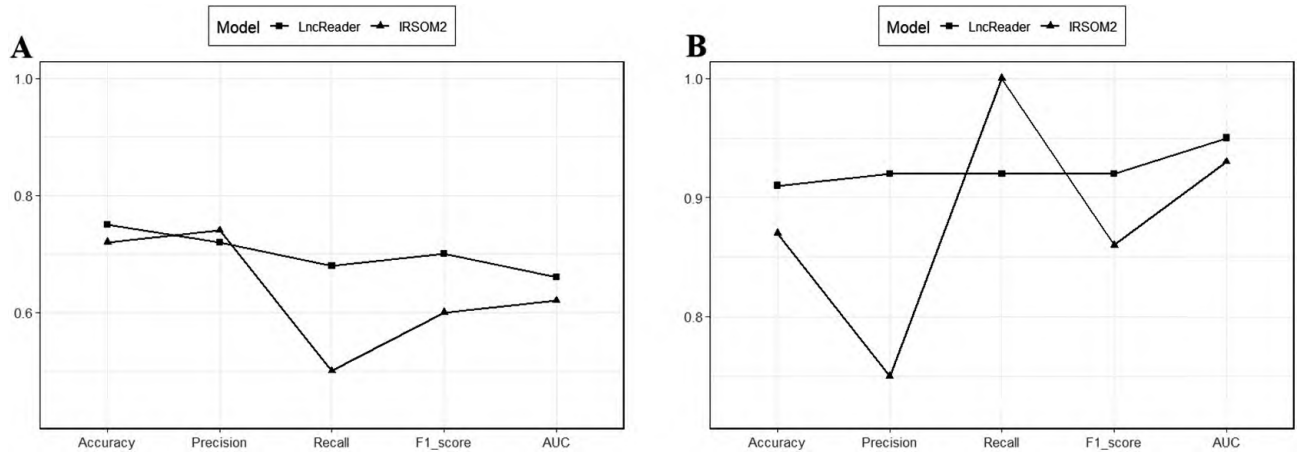

图2 LncReader与IRSOM2在拟南芥（A）与玉米（B）数据集上的五项分类指标比较

表3 LncReader与 IRSOM2在拟南芥和玉米数据集上的分类性能评估指标

模型名称	物种	准确率	精确率	召回率	F1分数	AUC
LncReader	*Arabidopsis thaliana*	0.76	0.71	0.70	0.70	0.66
IRSOM2	*Arabidopsis thaliana*	0.69	0.75	0.45	0.56	0.59
LncReader	*Zea mays*	0.75	0.72	0.68	0.70	0.66
IRSOM2	*Zea mays*	0.72	0.74	0.50	0.60	0.62

在拟南芥数据集中，LncReader在准确率（0.76）、召回率（0.70）和F1分数（0.70）方面均高于IRSOM2（准确率0.69、召回率0.45、F1分数0.56），体现出较为均衡的分类性能。虽然IRSOM2在精确率上略高（0.75），但由于其召回率偏低，导致整体F1分数明显下降，提示其对正类样本的识别存在不足，可能出现较多的漏判现象。

在玉米数据集中，两种模型的表现差距有所缩小，但LncReader在多个指标上仍略占优势。其中，LncReader的F1分数为0.70，高于IRSOM2的0.60，说明其在保持一定精度的同时具有更好的召回能力。与拟南芥一致，IRSOM2在精确率上略高（0.74），但在召回能力上的劣势同样明显（0.50）。

整体来看，LncReader在两个物种中均表现出更高的分类稳定性和更好的均衡性，尤其在召回率与F1分数上相对占优，更适用于对正类识别能力要求较高的植物tr-lncRNA筛查任务。而IRSOM2的性能更偏向于精度优先，其较高的精确率可能适用于高置信度小规模识别场景，但在覆盖性和灵敏性方面存在明显局限。

此外，结合两个物种的整体表现可以看出，两个模型在植物tr-lncRNA识别任务中的总体性能均未达到理想状态。模型间差异在统计层面虽存在，但不具显著断层，这进一步验证了前述ROC分析所得：现有模型在植物系统中的识别能力整体偏弱，其结构特征设计与训练策略在植物适应性上仍有较大优化空间。

3.3 模型适用性分析与小结

从拟南芥和玉米两个植物物种的评估结果来看，LncReader和IRSOM2均具备一定的tr-lncRNA判别能力，但整体性能尚处于中等偏低水平。模型间的差异主要体现在分类策略与特征依赖性上，其适用场景也有所不同。

LncReader在召回率与F1分数上的表现相对更优，具有较好的分类均衡性，在两个物种中都保持了

稳定的性能。其深度学习结构能够综合序列上下文、结构特征和翻译信号，对结构复杂或信号非典型的tr-lncRNA保持一定的敏感性，适合用于大规模初步筛选任务，尤其适用于研究者关注“尽可能发现潜在正类”的场景。

相比之下，IRSOM2依赖于传统的ORF统计特征与核苷酸频率建模，虽然在精确率上略具优势，但召回能力较低，F1分数相对较差。其分类更为保守，在实际应用中更适合用于小规模验证性研究或辅助分类阶段，尤其适用于对分类精确率有较高要求但可接受较多漏判的情境。

值得注意的是，尽管LncReader在各项指标中略优于IRSOM2，二者的性能差距总体并不显著，且均未在植物系统中表现出高度鲁棒性或泛化能力。这一结果表明，当前基于动物数据开发的主流预测模型在迁移至植物tr-lncRNA识别任务时，面临实际表现下降的问题。

造成上述局限的可能原因包括：植物lncRNA的序列保守性差、结构异质性强，以及小肽翻译信号的不典型性，使得传统以ORF规则或深度语义建模为核心的预测框架无法有效捕捉植物系统中的tr-lncRNA特征模式。此外，训练样本的物种偏倚亦可能限制模型在植物上的泛化表现。

综上，两种模型在植物tr-lncRNA识别中均表现出一定适用性，但存在不同的优势侧重与使用边界。现有方法尚不能完全满足植物系统中对tr-lncRNA识别的准确性与通用性要求，亟需开发基于植物组学数据训练的定制化模型，或引入迁移学习、弱监督标注等策略提升跨物种预测能力。

4 结束语

本文基于拟南芥与玉米两个植物物种构建高置信度tr-lncRNA数据集，对LncReader与IRSOM2两种主流tr-lncRNA预测模型的识别性能进行了系统比较。两种模型在植物系统中均具备一定的判别能力，但整体分类性能相对有限，且存在跨物种迁移后的识别能力衰减问题。其中，LncReader在召回率与F1分数方面略优于IRSOM2，而IRSOM2在精确率上表现更为保守。实验结果表明，当前主流模型尚未充分适应植物lncRNA在结构复杂性、序列保守性差及翻译信号不典型性等方面的特点，提示植物系统下tr-lncRNA的识别仍需结合植物特异性组学特征，开发更加适配的建模策略。本文结果为后续构建植物专属tr-lncRNA预测框架提供了基础参考。

参考文献(References):

[1] Ruiz-Orera J, Messeguer X, Subirana J A, et al. Long non-coding RNAs as a source of new peptides[J]. elife,2014,3:e03523.

[2] Sruthi K B, Menon A, Vasudevan Soniya E. Pervasive translation of small open reading frames in plant long non-coding RNAs[J]. Frontiers in Plant Science, 2022, 13:975938.

[3] Röhrig H, John M, Schmidt J. Modification of soybean sucrose synthase by S-thiolation with ENOD40 peptide A[J]. Biochemical and biophysical research communications,2004,325(3):864-870.

[4] Fesenko I, Kirov I, Kniazev A, et al. Distinct types of short open reading frames are translated in plant cells[J]. Genome research,2019,29(9):1464-1477.

[5] WANG S, TIAN L, LIU H, et al. Large-scale discovery of non-conventional peptides in maize and Arabidopsis through an integrated peptidogenomic pipeline[J]. Molecular plant,2020,13(7):1078-1093.

[6] Chilley P M, Casson S A, Tarkowski P, et al. The POLARIS peptide of Arabidopsis regulates auxin transport and root growth via effects on ethylene signaling[J]. The Plant Cell,2006,18(11):3058-3072.

[7] LIN X, LIN W, KU Y S, et al. Analysis of soybean long non-coding RNAs reveals a subset of small peptide-coding transcripts[J]. Plant physiology,2020,182(3):1359-1374.

[8] LIU T, ZOU B, HE M, et al. LncReader: identification of dual functional long noncoding RNAs using a multi-head self-attention mechanism[J]. Briefings in Bioinformatics, 2023,24(1):bbac579.

[9] Postic G, Tav C, Platon L, et al. IRSOM2: a web server for predicting bifunctional RNAs[J]. Nucleic Acids Research, 2023,51(W1):W281-W288.

[10] CHEN Y, LI D, FAN W, et al. PsORF: a database of small ORFs in plants[J]. Plant Biotechnology Journal, 2020, 18(11):2158.

[11] Ali U, TIAN L, TANG R, et al. A comprehensive atlas of endogenous peptides in maize[J]. iMeta, 2024(11): e247.

[12] WANG S, TIAN L, LIU H, et al. Large-scale discovery of non-conventional peptides in maize and Arabidopsis through an integrated peptidogenomic pipeline[J]. Molecular plant,2020,13(7):1078-1093.

[13] LI F, ZHENG Q, Vandivier L E, et al. Regulatory impact of RNA secondary structure across the Arabidopsis tran-

scriptome[J].The Plant Cell,2012,24(11):4346–4359.
[14] Hsu P Y, Calviello L, WU H Y L, et al. Super–resolution ribosome profiling reveals unannotated translation events in Arabidopsis[J]. Proceedings of the National Academy of Sciences,2016,113(45):E7126–E7135.
[15] Lukoszek R, Feist P, Ignatova Z. Insights into the adaptive response of Arabidopsis thaliana to prolonged thermal stress by ribosomal profiling and RNA–Seq[J]. BMC plant biology,2016,16:1–13.
[16] XU G, GREENE G H, YOO H, et al. Global translational reprogramming is a fundamental layer of immune regulation in plants[J].Nature,2017,545(7655):487–490.
[17] Gawro ń ski P, Jensen P E, Karpi ń ski S, et al. Pausing of chloroplast ribosomes is induced by multiple features and is linked to the assembly of photosynthetic complexes[J].Plant Physiology,2018,176(3):2557–2569.
[18] Sotta N, Chiba Y, Miwa K, et al. Global analysis of boron–induced ribosome stalling reveals its effects on translation termination and unique regulation by AUG–stops in Arabidopsis shoots[J]. The Plant Journal,2021,106(5):1455–1467.
[19] WU H Y L, HSU P Y. A custom library construction method for super–resolution ribosome profiling in Arabidopsis[J]. Plant Methods,2022,18(1):115.
[20] LEI L, SHI J, CHEN J, et al. Ribosome profiling reveals dynamic translational landscape in maize seedlings under drought stress[J]. The Plant Journal,2015,84(6):1206–1218.
[21] Chotewutmontri P, Barkan A. Dynamics of chloroplast translation during chloroplast differentiation in maize[J]. PLoS genetics,2016,12(7):e1006106.
[22] Chotewutmontri P, Barkan A. Multilevel effects of light on ribosome dynamics in chloroplasts program genome–wide and psbA–specific changes in translation[J]. PLoS genetics,2018,14(8):e1007555.
[23] JIANG J, CHAI X, Manavski N, et al. An RNA chaperone–like protein plays critical roles in chloroplast mRNA stability and translation in Arabidopsis and maize[J]. The Plant Cell,2019,31(6):1308–1327.

（上接第32页）

编辑及隐私保护提供了解决方案，未来可应用于需符合隐私法规的场景。例如，某医院将患者诊疗记录上链，包含诊断结果（公开）和身份证号（敏感），患者请求删除身份证号后，节点仅移除对应数据段，保留诊断结果的Merkle证明供第三方验证；在供应链溯源的商品物流信息中包含供应商机密报价，供应商可选择性删除报价字段，保留物流时间戳的验证能力。由此，通过Merkle树，节点可删除特定数据条目而不影响整个交易的存在性，这对隐私保护至关重要。

参考文献(References):

[1] 沈海波，陈强，黄海．语义区块链研究综述[J]. 计算机应用研究，2021,38(7):1937–1942.
[2] 链上物联网．一文读懂：区块链中的Merkle树[EB/OL]. (2021–05–20) [2024–04–20]. https://blog. csdn. net/XYlittlework/article/details/117061970.
[3] 王嘉瑶，王婷，袁文亮，等．分布式账本技术的发展历程研究综述[J]. 计算机应用研究，2023,40(3):641–648.
[4] SONG H, WEI Y, QU Z, et al. Unveiling Decentralization: A Comprehensive Review of Technologies, Comparison, Challenges in Bitcoin, Ethereum, and Solana Blockchain[EB/OL]. arXiv, 2024. [2024–04–20]. https://arxiv.org/abs/2404.00007.
[5] Merkle R C. A Digital Signature Based on a Conventional Encryption Function[C]//Conference on the theory and application of cryptographic techniques. Berlin: Springer–Verlag,1987:369–378.
[6] 田海博，何杰杰，付利青．基于公开区块链的隐私保护公平合同签署协议[J]. 密码学报，2017,4(2):187–198.
[7] 王化群，吴涛．区块链中的密码学技术[J]. 南京邮电大学学报(自然科学版)，2017,37(6):61–67.
[8] 黄根，郜一波，徐云．区块链中Merkle树性能研究[J]. 计算机系统应用，2020,29(9):237–243.
[9] Davies J. Enhanced scalability and privacy for blockchain data using Merklized transactions[J]. Frontiers in Blockchain,2024(6).
[10] 吴梦宇，朱国胜，吴善超．基于Merkle树的区块链数据修改方法研究[J]. 信息通信，2020(10):10–12,16.